Introduction

Shannon Pileggi

STAT 217

OUTLINE

A study of 120 articles from 1998 to 2006 in 10 leading international psychology journals found that statistical inference was used in about _____ of articles.

1. 5%
2. 30%
3. 50%
4. 70%
5. 95%

About Dr. Pileggi

Degrees

- BS Mathematics and Hispanic Studies
- MS Biostatistics
- PhD Biostatistics

Personal

- Married
- Have a 2 year old daughter
- Have 2 dogs
- Enjoy: bike commuting, soccer, disc golf, hiking, board games

Course Philosophy

Most assignments are due **before** we cover a topic.

- Pre-lab assignemnts
- Topic readiness

Then in class we will discuss elements of your assignment and go deeper into the material.

- This gives you *repetition*.
- This gives you an opportunity to ask more and better questions in class.
- This allows for you get a deeper understanding of the material during your face to face time with the instructor.

Lab Groups vs Project Teams

Statistics is a team sport!

Lab Groups

- Size: 3-4
- Lab groups will be randomly assigned by the instructor
- Each of the three course units will have a different group assignment
- In class, sit with your lab group to discuss the group exercises

Project Team

- Size: 3-4
- You may select your project team
- Project team stays constant throughout the quarter

Statistics in the News

- Psychology - Facebook use 'undermines well-being'
- Linguistics - English language 'originated in Turkey'
- Education - US college degree still worth it, says study
- Sociology - Disparities: CT Scans More Likely for White Children
- Public Health - Young cannabis smokers run risk of lower IQ, report claims
- Medicine - Alzheimer's disease drug shelved after trial failure
- Neuroscience - Obesity hastens cognitive decline
- Anthropology - Testicle size 'link to father role'
- Sports - Does it make statistical sense to sack a football manager?
- Gambling - PhD wins Lottery 4 Times

What is statistics?

Statistics is way to make sense of data. It deals with

- data collection
- data analysis
- interpretation of results

Generally, we use statistics to make decisions about a population based on information obtained from a randomly selected *sample*.

Getting Started

Foundations

Evidence in Statistics

Describing variables

Example research question

What percent of US households earn over $200,000 annually?

Anecdotal evidence vs statistical evidence

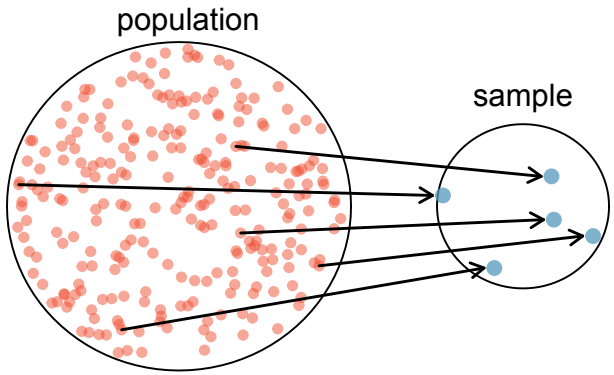
Informal observations constitute **anecdotal** evidence.

- Anecdotal evidence may be true, but is often based on small samples that are not representative of an entire population of interest.
- We would need a more formal study to make strong conclusions with statistical evidence about relationships observed or findings presented.

Population vs Sample

To answer a research question, you identify the **population** of interest from which you will collect your **sample** data.

- A **population** is the set of all subjects of interest.
- A **sample** is the subset of the population of interest on which you collect data.



Group Exercise

Collecting data on everyone is called a **census**.

Suppose I wanted to know about the income level of United States citizens. Why don't we just collect data on everyone in the population, ie, perform a census?

Example research - helper vs hinder

- *Nature* 2007 (Hamlin, Wynn, Bloom) Social evaluation by preverbal infants
- Research question: Is a moral compass innate or is it learned?
- 16 10-month-old infants from New Haven, CT witness a "climber" trying to make it up a hill
- the climber then faces either a "helper" (climber pushed to the top) or a "hinderer" (climber pushed to the bottom)
- the child was encourage to select either the helper or the hinderer to play with

See videos from the *Nature* article.

Group Exercise

What is the:

1. sample?

2. population?

Description vs Inference

After you have collected data, you **describe** the characteristics of your sample. You use the data collected from the sample to make **inference** on the population of interest.

- **Descriptive statistics** are used to summarize the collected data through numbers such as averages and percentages.
- **Inferential statistics** are used to draw conclusions about a *population*, based on the data obtained from a *sample* of the population.

Group Exercise

The researchers find that 14 out of 16 (87.5%) infants prefer the helper toy, and conclude that 63-98% of infants in general would select the helper toy.

Which parts refer to descriptive vs inferential statistics?

1. both statements are descriptive statistics
2. both statements are inferential statistics
3. *descriptive* = 87.5% of infants prefer the helper toy;
   inferential = 63-98% of infants in general would select the helper toy
4. *descriptive* = 63-98% of infants in general would select the helper toy;
   inferential = 87.5% of infants prefer the helper toy

Parameters vs statistics

A **parameter** is a numerical summary of the *population*.

- We want to make inference on parameters
- The true value of a parameter is unknown
- We denote parameters with Greek letters

A **statistic** is a numerical summary of the *sample*.

- We calculate statistics from our *sample* data.
- Statistics are our best estimate of parameters.
- We denote statistics with lower case letters, bars, and hats

Parameters:

population mean	μ
population standard deviation	σ
population proportion	p

Statistics:

sample mean	$\bar{x}$
sample standard deviation	s
sample proportion	$\hat{p}$

Group Exercise

The researchers find that 14 out of 16 (87.5%) infants prefer the helper toy.

What is the parameter and the statistic?

1. Parameter = 87.5%, Statistic = 87.5%
2. Parameter = unknown, Statistic = unknown
3. Parameter = 87.5%, Statistic = unknown
4. Parameter = unknown, Statistic = 87.5%

Summary

- Sample data are an approximate (imperfect) reflection of the population data.
- What is see in the data is not exactly as things are in the population.
- Statistical inference is about describing what you think is likely to be happening in the population, based on the observed sample data.
- In order to do this, we need to understand *variation* in our data.

Group Exercise

Think about the Helper vs Hinderer study... What are two possible explanations for why 14 out of 16 infants selected the helper toy?

1.

2.

Which explanation do you think is more plausible?

Statistical logic

With statistical reasoning, we generally consider two possible models:

1. The data arose from random chance
2. The data didn't arise from random chance - something is really going on here

Then we come up with evidence to differentiate between the two models.

Simulating a chance model

- coin flip = an individual infant's selection
- heads = represents infant selecting helper
- tails = infant selects hinderer
- chance of heads = 0.5, the probability that the infant randomly selects the helper
- one repetition = one set of 16 coin flips to represent the 16 infants in the study

Plot of class results for number of heads out of 16 coin flips

Group Exercise

Analyzing the evidence:

- Would it be surprising to have 14 out of 16 infants select the helper just by chance?

- Does this *prove* or *provide evidence* that infants have a moral compass?

Getting Started

Foundations

Evidence in Statistics

Describing variables

Data sets: Selling Mario Kart on eBay

Columns indicate **variables**

↓

Rows indicate **observations** →

obs	nBids	cond	startPr	totalPr	shipSp	wheels
1	20	new	0.99	51.55	standard	1
2	13	used	0.99	37.04	firstClass	1
3	16	new	0.99	45.50	firstClass	1
4	18	new	0.99	44.00	standard	1
5	20	new	0.01	71.00	media	2
6	19	new	0.99	45.00	standard	0
7	13	used	0.01	37.02	standard	0
8	15	new	1.00	53.99	upsGround	2
9	29	used	0.99	47.00	priority	1
10	8	used	19.99	50.00	firstClass	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
143	13	new	1.00	54.51	upsGround	2

Variables

A **variable** is any characteristic observed in a study that is measured on a **unit of observation**.

- A variable is called **categorical** if each observation belongs to one of a set of categories. Categorical variables typically contain descriptive words or phrases.
- A variable is called **quantitative** if observations take on numeric values.

Which variables are categorical and which are quantitative?

obs	nBids	cond	startPr	totalPr	shipSp	wheels
1	20	new	0.99	51.55	standard	1
2	13	used	0.99	37.04	firstClass	1
3	16	new	0.99	45.50	firstClass	1
4	18	new	0.99	44.00	standard	1
5	20	new	0.01	71.00	media	2
6	19	new	0.99	45.00	standard	0
7	13	used	0.01	37.02	standard	0
8	15	new	1.00	53.99	upsGround	2
9	29	used	0.99	47.00	priority	1
10	8	used	19.99	50.00	firstClass	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
143	13	new	1.00	54.51	upsGround	2

Group Exercise

In an article published in the *British Medical Journal* (2004), researchers reported that heart transplantations at St. George's Hospital in London had been suspended in September 2000 after a sudden spike in mortality rate. Of the last 10 heart transplants, 80% had resulted in deaths within 30 days of the transplant.

What is the unit of observation?

1. St. George's Hospital
2. London
3. a heart transplant
4. 80% resulted in deaths
5. the 30 days studied
6. whether or not the patient died

Categorical variables

There are also more ways to describe categorical variables.

- A categorical variable that only has two categories is said to be **dichotomous** (e.g., students that live on or off campus).
- There is not a special term for categorical variables with more than two categories (e.g., religious beliefs can be classified as Christian, Muslim, Hindu, etc.)
- A categorical variable is said to be **ordinal** if its categories have a natural ordering. (e.g., year in college can be classified as freshman, sophomore, junior, senior)

Group Exercise

In the helper vs hinder study, where we found that that 14 out of 16 (87.5%) infants prefer the helper toy, what was...

1. the unit of observation?
2. the variable studied?
3. the parameter?
4. the statistic?